DOCUMENT RESUME

ED 373 555 FL 022 355

AUTHOR	Geranpayeh, Ardeshir
TITLE	Are Score Comparisons across Language Proficiency Test Batteries Justified?: An IELTS-TOEFL Comparability Study.
PUB DATE	94
NOTE	18p.; For serial publication in which this paper appears, see FL 022 351.
PUB TYPE	Reports - Research/Technical (143) -- Journal Articles (080)
JOURNAL CIT	Edinburgh Working Papers in Applied Linguistics; n5 p50-65 1994
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Comparative Analysis; *English (Second Language); Foreign Countries; *Graduate Students; Higher Education; *Language Proficiency; *Language Tests; Scores; Student Attitudes; Test Format; *Test Validity
IDENTIFIERS	*International English Language Testing System; Test Equivalence; *Test of English as a Foreign Language

ABSTRACT

This paper reports on a study conducted to determine if comparisons between scores on the Test of English as a Foreign Language (TOEFL) and the International English Language Testing Service (IELTS) are justifiable. The test scores of 216 Iranian graduate students who took the TOEFL and IELTS, as well as the Iranian Ministry of Culture and Higher Education Test of English Proficiency (MCHE), from 1990-92 were compared. The study found high to moderate correlations between TOEFL and IELTS scores. Comparisons indicate that a score of 6 on IELTS is roughly equated with 600 on TOEFL, the minimum requirement for non-native speakers to gain admittance to most English-language graduate schools. A score of 6.5 on IELTS is roughly equated with 600 on TOEFL, the minimum requirement for non-native speakers to gain admittance to a linguistics department in most English-language graduate schools. The scores of the most proficient subjects on the two tests were found to be less comparable than the scores of less proficient subjects. An appendix contains a correlational matrix among subsections of the three tests. (Contains 23 references.) (MDM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 373 555

Are Score Comparisons Across Language Proficiency Test Batteries Justified? An IELTS-TOEFL Comparability Study.

Ardeshir Geranpayeh (DAL)

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Brian Parkinson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL 022 355

BEST COPY AVAILABLE

ARE SCORE COMPARISONS ACROSS LANGUAGE PROFICIENCY TEST BATTERIES JUSTIFIED?: AN IELTS - TOEFL COMPARABILITY STUDY.

Ardeshir Geranpayeh (DAL)

Abstract

Many academic institutions in the UK and Australia require their non-native candidates to provide a proof of a certain band score on IELTS or its equivalent score on TOEFL as evidence of English proficiency to pursue a course of study. This study is concerned with whether score comparisons across TOEFL and IELTS are justified. The results reported here suggest that score comparisons across TOEFL and IELTS are possible but institutions should be cautioned about the comparability of the test scores and should allow for possible extraneous factors affecting these scores.

1. Introduction

1.1 Introduction

Hundreds of thousands of individuals throughout the world take various English language proficiency tests each year to demonstrate their proficiency in English as a foreign language. The scores of such tests will then be used by different institutions for screening their candidates for a number of different purposes such as offering employment, advancement in a career, or admission to an educational programme. In most cases, the selection of candidates is affected by the results of these tests. Thus, any variability in the scores of such tests might affect job opportunities or perhaps life chances of individuals. This makes the interpretation of the scores an extremely heavy responsibility.

Test scores could be related to various aspects of proficiency demonstrating the candidates' language ability in different skills, i.e. writing, reading, speaking, or listening in a given language. In the last three decades, numerous methods and test batteries have been developed to measure different aspects of language proficiency of non-native speakers. Depending on the nature of the test population and the purposes to which the test scores are put, the tests presumably differ from one another. In most cases, differences in methods and purposes are considered as evidence for the incomparability of LP tests. Yet, where the statistical evidence is concerned, the tests are validated against one another and their results are compared to show the degree of similarity between the traits they are measuring.

On the other hand, academic institutions are interested only in a clear cut-off point score of, say, 600 on TOEFL or its equivalent 6.5 on IELTS as evidence of their non-native speakers' suitability to pursue a course of study. What does it mean to have a specific score on a test? How can a quantitative value obtained in an hour's

testing period predict the future success of a candidate in following a career? How can different scores obtained in different batteries be equated to one another? These and many more questions have been raised in the literature of language proficiency testing.

This paper is an attempt to clarify one of the relevant issues in comparability studies, that is, whether score comparison across test batteries is justified. The paper is limited in scope to the study of two influential LP tests currently administered worldwide: TOEFL and IELTS. We will begin by pointing to the differences in British and North American traditions in language testing. This is followed by a brief review of the effect of the test methods on the measures of a construct. Then, the question of the research is discussed. In the method section, reviews of the tests concerned here will provide the basis for score comparisons across test batteries. Results of the comparisons will be reported and discussed in detail. Finally, the conclusion will sum up the discussions.

1.2 Test methods

There is a general belief that British and North American EFL proficiency tests represent radically different approaches to language test development. North American tradition in language testing is heavily based on psychometric properties of tests. Issues such as reliability and concurrent and predictive validity are of particular interest in this tradition. Hence, objectivity of scoring and generalisibility of the results play a dominant role in the development of test methods. For example, multiple-choice items are [illegible] used in testing receptive skills to gain desired internal consistency, even if the tests [illegible] expected to measure communicative competence as is the case in Functional Testing (Farhady 1980). Moreover, in order to achieve high inter-rater reliability, the use of trained scorers and detailed specific instructions in conducting an interview are highly recommended for testing productive skills in this tradition.

When we examine the British tradition, it is observable that the emphasis is on the specification of test content and expert judgement. While reliability (degree of generalisibility of the results) receives less attention in this tradition, content and face validity are the major concerns of the test designers. It may follow that British tests enjoy more variability in their formats and include various communicative activities.

Different test methods might well affect the performance of the candidates taking the tests. The characteristics of test methods which influence test performance have long been studied by many researchers in language testing. Research has shown that test performance varies as a function both of an individual's language ability and of the characteristics of test methods. Some test takers, for example, might perform better in the context of a laboratory speaking to a microphone than they would in front of a panel of judges in an oral interview. Some test takers might find it easier to choose responses from among alternatives in a multiple-choice test of vocabulary than to complete an open-ended cloze format of a similar test. Completion of isolated sentences as opposed to completion of blanks in a text, live versus recorded speech, aural in contrast to written tests, are but a few examples of how methods of testing may vary. These characteristics of test methods may, in turn, influence the test performance, casting doubt on the reliability and validity of language tests. Controlling these characteristics thus becomes an important issue in the theory and

practice of language testing.

The study of test methods dates back to 1959 when Campbell and Fiske (1959) showed that method variance might influence the measures of a construct. They argued that a hypothetical large correlation between two traits, let us say A and B, and no correlation between traits A and C,might be a function of method variance common to the measures A and B and not to C, if the measures A and B are obtained by one method and that of C by another method. To control the method effect, they proposed a multitrait-multimethod (MTMM) design for validating tests. The main focus of the MTMM design is to separate trait and method factors. It recognizes that "any test score is a function of both the trait it intends to measure and of the method by which it is measured" (Bachman and Palmer 1979:54). Therefore, the method involved in measuring might become as important as the trait it is intended to measure.

According to MTMM design, to observe the validity of a test, that is, to see whether the test is measuring what it purports to test, the application of more than one method seems necessary. If independent methods testing the same construct do tend to correlate highly, it is concluded that convergent validity is achieved. On the other hand, to achieve discriminant validity, i.e. to show that there are independent traits irrespective of the methods applied, introduction of more than one trait in the analysis is necessary. Low correlation between different traits indicates that they are really different from one another and hence discriminant validity is achieved.

As it stands, independence of methods is an important issue in validity as well as reliability studies. Convergence of independent methods claiming to test similar constructs is a proof of the validity of a test. However, in the case of reliability, convergence of similar methods is indicative of the reliability of the test. Since independence is a matter of degree, it may be concluded that reliability and validity can be considered to be on a continuum, depending on the degree of independence of test methods. That is,

> "Reliability is the agreement between two efforts to measure the same trait through maximally similar methods. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods." (Campbell and Fiske 1959: 83)

The MTMM design of Campbell and Fiske was influential for those interested to know whether the techniques testers use distort the results that they obtain. Bachman and Palmer (1981), for example, used a complex MTMM research design to investigate the comparative influences of two traits (speaking and reading) and three methods (interview, translation and self-rating). They found that scores from self-ratings loa'ed consistently more highly on method factors than on specific trait factors, and that translation and interview measures of reading loaded more heavily on method than on trait factors. Similar results were obtained in another study by the same researchers. Bachman and Palmer (1982) found that scores from both self-ratings and oral interviews consistently loaded more heavily on test method factors than on specific trait factors, while the scores from the multiple-choice and writing tests were least affected by method factors. A number of other studies have also examined the effect of test methods on test performance (see Alderson 1978, Bachman 1982, Lewkowicz 1983, Shohamy 1984, Chappelle and Abraham 1990).

What are the characteristics of test methods? The facets of test methods can be viewed from different perspectives. Bachman (1990:119) proposes a comprehensive framework for studying the facets of test methods. His framework comprises five main categories: facets of the testing environment, facets of the test rubric, facets of the input, facets of the expected response, and relationship between input and response. The large number of dimensions along which test methods vary in Bachman's framework are reflections of the variety of testing techniques that are used in language tests, and the ways in which these techniques vary.

Bachman's framework has been used for examining the various dimensions or facets of test methods in a large scale study, namely the Cambridge - TOEFL Comparability Study (Bachman, Davidson, and Foulkes 1993). This study offers an interesting suggestion: that different methods as diverse as Cambridge and ETS test batteries not only tap, to a large degree, similar abilities of the subjects in the sample concerned but also measure these abilities in much the same way. Among the findings of this study is the legitimacy of score comparisons across these two test batteries.

1.3 Scope of the present study

Bachman et al. (1993) suggested that score comparison across ETS tests and UCLES tests (CPE) could be made in a meaningful way. This would mean that institutional administrators across the Adantic need not require separate test results for individuals who have already taken one of the test batteries. This will save time and money both for the individuals taking the tests and for the institutions offering the opportunities (admission, jobs etc.). If it is the case that score comparison is legitimate across Cambridge proficiency tests and ETS tests, the same comparison should also be possible between ETS tests (namely TOEFL) and IELTS (designed by UCLES). In addition, most universities in Australia and the UK require their non-native graduate candidates to provide a score on either TOEFL or IELTS as a proof of their proficiency in English. It seems that these institutions are practically equating the scores from TOEFL with those of IELTS.

In this research we are looking for the justification of score comparisons across TOEFL and IELTS. So the following questions are raised. Are TOEFL and IELTS comparable? Is there any consistent relationship between TOEFL and IELTS scores across time? Do preparation courses affect the performance of subjects in LP tests?

This study is also limited in scope to the study of Iranian graduate students' scores on TOEFL and IELTS between 1990 and 1992. Iranian graduate students who are intending to continue their studies by taking a PhD degree in English speaking countries are required to sit either TOEFL or IELTS. In many cases they sit both tests. The Ministry of Culture and Higher Education (MCHE) in Iran has developed a TOEFL-like test (MCHE) for screening the candidates before sitting the above tests. Only those who score above 50 (0-100 scale) on MCHE will be allowed to sit TOEFL or IELTS. The data presented here are based on the scores of those candidates who have sat all the three tests (IELTS, TOEFL, and MCHE) during 1990-1992.

BEST COPY AVAILABLE

2. Method

2.1 Reviews of proficiency tests

Prior to any discussion, analyses of the characteristics, activities and score bands as well as the underlying constructs of each test seem to be warranted.

2.1.1 TOEFL

The Test of English as a Foreign Language (TOEFL), a highly secure test, is the most widely administered, standardised , multiple-choice test of language proficiency (1963-1994). TOEFL is administered 12 times a year, a new equated form each month, at more than 1,100 centres in 170 countries and areas and its results are used by some 2500 universities and colleges in the U.S, Canada and other countries for a variety of academic subject areas. According to ETS (1992) some 1,178,193 students seeking admission to institutions in the United States or Canada took the test from July 1989 to June 1991. The test is designed to 'evaluate English LP of individuals whose native language is not English, most often those wishing to study in North American universities and colleges' (Stevenson 1987:79), it is recommended for students at 11th grade level or above.

2.1.1.1 The structure of TOEFL

The test comprises three sections (since 1976), each separately timed: Listening Comprehension (50 minutes), Structure and Written Expression (25 minutes), and Reading Comprehension and Vocabulary (60 minutes). All the items are in 4-MC format. TOEFL total scores range between 227-677 without any pass/fail scores. Nevertheless, institutions require different ranges of scores for different subject areas.

The TOEFL is, without a doubt, the most reliable as well as the most researched of all foreign LP tests, having been under constant revision and empirical research study for the past thirty years. The TOEFL Research series as of Summer 1993, consisted of 45 Research Reports and 6 Technical Reports. Over the years, TOEFL has been used as a criterion for the validation of other tests. Among the most recent attempts of this kind is the Cambridge - TOEFL Comparability Study (Bachman, et al. forthcoming).

2.1.1.2 Reliability and Validity

The reliability of the test has repeatedly been reported to be satisfactory. Stevenson (1978) reports that ' the average reliabilities for 12 forms (administered in 1981-1982) are 0.89, 0.87, and 0.89 for the three sections, and 0.95 for the total score' (1987: 80). This is well within the desirable range for this type of test.

Validity of a test, by definition, depends on the extent to which a test measures what it purports to measure. TOEFL is intended to measure the English-language proficiency of non-native speakers of English who wish to study in North American universities. Hence, the content of the test should be representative of the social situations to which the examinees are expected to be exposed. The specification of such a context is not an easy task, given the wide range of TOEFL populations and target language-use situations. It seems that the traditional techniques of contrastive

analysis and error analysis are not appropriate for content selection of TOEFL. Like all proficiency measures, the content validity of TOEFL depends on the degree to which experts perceive it to be valid. Stevenson points out that:

> 'TOEFL does agree that content is best specified by experts, and does rotate membership in this group often to avoid stagnation or the dominance of one view, and leads to the reasonable conclusion, if not demonstration, that the content of TOEFL in general, is representative.' (1987: 81)

As for the construct validity of the test, we know that construct validity concerns 'the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of abilities, or constructs' (Bachman 1990: 255). The abilities involved in the construct of LP are theoretical, yet to be defined and agreed upon. Hence they constrain our efforts to test the extent to which we can make inferences about these hypothetical abilities on the basis of test performance. Unless we have a clear definition of the construct, we cannot claim to have measured it. TOEFL constructors seem to be very conservative in stating what construct they purport to test. For example, the TOEFL Bulletin of Information for TOEFL TWE and TSE, 1992-1993 (ETS 1992: 3) states that the Vocabulary and Reading Comprehension section of the test 'measures ability to understand non-technical reading matter' in standard written English. It goes on to talk about the multiple-choice format of the questions implied, stated or otherwise. But it never explicitly defines the construct. As Pearce (1992:668) pinpoints, 'the construct of reading that is measured in the TOEFL reading test is not made explicit in the ETS literature'. Indeed ETS cannot make it explicit as there is no promising definition in the state of the art at present. Having said all this, there seems to be a general agreement in ETS that there exists a general proficiency factor which is divisible by skills and components.

2.1.2 IELTS

The International English Language Testing Service (IELTS) has been developed (1980-1994) jointly by the British Council and the University of Cambridge Local Examinations Syndicate (UCLES) to determine whether students' ability in English would meet the demands of a course of study in Britain and Australia. The early versions of the test (ELTS 1980-1989) comprised 6 subject specific areas in addition to a general section. The test reflects the ideas of communicative language teaching and is probably the first standardised communicative language test administered over a large population across the world. Some 37,455 non-native speakers of English are reported to have taken the test between 1981-1985 (Criper and Davies 1988). IELTS has been widely welcomed by the British and Australian universities as it claims to be a test of English for Specific Purposes (ESP).

Though the test was meant to be one of ESP, the final form includes an additional general section. The test follows the Munby (1978) communicative syllabus design. Carroll (1978) guided the test specifications on the basis of needs analysis. The analysis suggested a number of specific tests for different subject areas. However, in practice, large compromises and reductions were made, limiting the specific areas to six (from 1980 to 1989) and to three (since 1989), and perhaps only to one (from 1994), these changes being mainly determined by the British Council and UCLES, not by the students' needs.

2.1.2.1 The structure of IELTS

IELTS consists of two sections: General (G) and Modular (M). The general section consists of a listening test and an oral interview intended to test the oral skills. The Modular section, on the other hand, is intended to test the written skills: reading and writing. The modules are limited to two forms: Modules A, B, and C, for academic audiences, and Module GT, for non-academic general training purposes.

The listening part consists of thirty-five multiple-choice test items accompanied by a tape in four sections: 1) choosing from diagrams; 2) listening to an interview; 3) replying to questions; and 4) listening to a seminar. The interview is conducted face to face, individually, usually with a time lapse from the written test. It consists of two parts: general questions, and questions about candidates future plans. The subject is then assigned to one of the bands (1-9).

The overall formats of the modules (M1= Reading) are the same. They all contain texts taken from books, journals, reports, etc., related to a specific subject area and involve testees in study skills necessary for academic studies, with the exception of the nonacademic area. There are all together 40 M-C test items in each module. The three academic modules are: Science and Technology, Life Science, and Arts and Social Sciences. Each student selects one module only. The Writing test has two questions in each module. The first question requires the testee to bring in his/her own experience and views on the basis of the reading texts. The second question is strictly limited to the information available in the text. Both tasks require the testees to write short paragraphs.

2.1.2.2 Reliability and Validity

There are no published statistics on IELTS except those reported by Alderson (1993) based on a trial test. Aside from the variations in the size of the trial population in different modules (not all students took every test in the battery), the reliabilities reported are acceptable. However, that of Module GT is questionable.

Table 1. Reliabilities Reported for IELTS. Trial Test. Alderson 1993

Tests	G1	MA	MB	MC	MGT	G2
Rel	0.86	0.90	0.91	0.88	0.79	0.87

G1= Grammar Test; MA= Science and Technology Reading Test; MB= Life Science Reading Test; MC= Arts and Social Sciences Reading Test; MGT= Nonacademic Reading Test; G2= Listening Test.

Alderson (1993) also reports the results of the reliabilities for the total test battery of listening, grammar, and reading tests ranging between 0.80-0.97, and that of the battery without the grammar test ranging between 0.76-0.96. Although the reliability of the total test battery declines in the absence of the grammar test, ' this decline is relatively unimportant, with the arguable exception of MGT, the General Training Model' (Alderson 1993: 215). The implication was that the grammar section should be dropped from the actual IELTS test. No reliability is reported for the total band

score.

A factor analysis of the test results reveals the emergence of a first dominant (general) factor followed by a second (writing) factor.

> 'In general, an analysis of reading, grammar, and listening yielded only one common factor. The addition of writing occasionally gave rise to a second factor.' (Ibid: 213)

Since Interview was not included in test analysis nor any other external criteria, it is difficult to predict what factors might have emerged had they been included in the analysis. The only statistics available in Alderson's (1993) report are the correlations between the two reading tests of the new (IELTS) test and the band score of the old ELTS subtests. The purpose of comparison 'was to enable the calculation of band scores to the test' (ibid: 214). There were significant variations in the relationship between the new and the old reading tests: readings MA correlated 0.39 while those of MC correlated 0.76.

The differences were justified on the assumption that the new IELTS test was an improvement on the old test and that the readings were not directly parallel to each other in content or topic.

Moderate correlations reported in the IELTS trial study between different modules support the ESP aspect of the test. IELTS does look and function like an ESP test. The test seems to be favoured more by its face validity than any other criteria. Due to the lack of published data, it is difficult to observe the extent to which the test measures what it purports to test. However, the factor analysis of the trial study does give evidence for the uni-factorial structure of the test.

IELTS seems to be based on a notion that proficiency is divisible by skill and as Alderson and Clapham (1992:164) report 'there are thus tests of the four macro-skills: reading, writing, listening, and speaking.'

2.1.3 MCHE

The Ministry of Culture and Higher Education Test of English Proficiency (MCHE) has been developed in Iran to assess the LP of Iranian graduate candidates who are awarded a scholarship to pursue their studies towards a PhD. At least three different versions of this test have been administered four times a year since 1989. The test comprises four multiple-choice sections: Listening Comprehension (30 items), Structure and Written Expression (30 items), Vocabulary (20 items), and Reading Comprehension (20 items). The total score is computed on the basis of the sum of the four sections (0-100). There is an additional writing (essay) section whose score (0-20) is reported separately. Due to administrative problems, the result of the latter section is not incorporated in this research.

There are no published data about the validity and reliability of this test. The structure of the test is very similar to that of TOEFL. The earliest version of MCHE was reported to have a correlation of 0.89 with TOEFL in 1989.

10
BEST COPY AVAILABLE

2.2 Subjects

The subjects were 1600 Iranian graduate students from different subject areas who sat for TOEFL and IELTS as well as for MCHE between 1990-1992. They were divided into two groups: Group A and Group B. Group A included students who sat for these tests between 1990 to early 1991 and for whom only the total scores for these tests were available. Group B included students who did the tests from early 1991 to mid 1992 and for whom both the total scores and the sub-section scores on each test were available. Only the scores of those who had done all the three tests were selected. Thus, only 113 and 103 subjects remained in Groups A and B respectively. Some students participated more than once in the tests. Only one score (the latest) of each student was counted for each test.

Moreover, most Group B subjects participated in TOEFL preparation courses during 1991-1992. Only a few participated in IELTS preparation courses. The IELTS sample materials, however, were distributed among all those from Group B who intended to sit for IELTS. The results reported here are based on 6 administrations of IELTS and 7 administrations of TOEFL.

3. Results

Relatively high correlations were found among Group A's scores on TOEFL and IELTS (table 2), while moderate correlations were found among Group B's scores on these tests.

Table 2: Correlations Between the Total Scores of the Tests: Group A Subjects

TESTS	TOEFL	IELTS
IELTS	0.8290	-----
MCHE	0.8339	0.7570

Table 3: Correlations Between the Total Scores of the Tests: Group B Subjects

TESTS	TOEFL	IELTS
IELTS	0.6671	-----
MCHE	0.6386	0.6072

By means of regression analyses score comparisons across tests were carried out. Tables 4 and 5 demonstrate the score comparisons across tests based on some of the key scores on MCHE.

Table 4: Score Comparisons Across Tests: Group A Subjects

TESTS	SCORES				
MCHE	50	60	70	80	90
IELTS	4.4	4.8	5.2	5.5	6
TOEFL	450	475	500	526	550

Table 5: Score Comparisons Across Tests: Group B Subjects

TESTS	SCORES				
MCHE	50	60	70	80	90
IELTS	4.6	5	5.3	5.7	6
TOEFL	460	495	530	565	600

The rest of the results relate to Group B subjects. Table 6 shows the mean score and standard deviation of the scores on each test.

Table 6: Descriptive Statistics

	Mean Score	Standard Deviation
MCHE	52	10
IELTS	4.7	0.7
TOEFL	468	54

A full correlational matrix of the relationships between the different subsections of the tests is given in Appendix I. A factor analysis was also conducted to find out the similarities between the two tests. Table 7 shows the results of the factor analysis. Varimax rotation extracted two factors. All the subtests of IELTS and TOEFL loaded mainly on the first general factor associated with general listening ability. The MCHE subtests loaded heavily on the second factor associated with general structure and reading comprehension.

Table 7: Factor Analysis: Rotated Factor Matrix

	FACTOR 1	FACTOR 2
MLC	47002	50915
MST	31146	63566
MVOC	25961	59368
MRC	10934	68662
IRC	68741	20749
IWR	62101	19941
ILC	69619	18214
ISP	49789	20632
TLC	75246	28392
TST	65528	49384
TRC	68478	44910

M = MCHE, I = IELTS, T = TOEFL, LC = Listening Comprehension,
ST = Structure, VOC = Vocabulary, RC = Reading Comprehension,
WR = Writing, SP = Speaking

Finally, to account for the effect of preparation courses (test effect), all the scores were converted to a scale of 0-20 so that the analysis of variance would become possible. A repeated-measures analysis of variance (MANOVA) was performed to find out whether there was any significant difference in the subjects' total score on the three different tests (TOEFL, IELTS, and MCHE). Table 8 illustrates the results of the MANOVA.

Table 8: Repeated-Measures Analysis of Variance

Source of Variation	SS	df	MS	F
Within Cells	355.82	204	1.74	36.95 *
Test	128.90	2	64.45	

* $p < 0.05$

The MANOVA detected a significant difference in the total test scores across the three batteries, suggesting the effect of the "test" factor. Of the three possible comparisons among the means, Tukey's WSD test shows that only the comparison between TOEFL and IELTS score was significant.

Table 9: Tukey Test of Differences Across Batteries

Test	TOEFL	MCHE	IELTS
	Mean = 10.73	Mean = 10.40	Mean = 9.22
TOEFL	--	0.33	1.51 *
MCHE	--	--	0.18
IELTS	--	--	--

* $p < 0.05$

4. Discussion

The reader should bear in mind that the intention of this research was not to carry out a full comparability study between IELTS and TOEFL. Rather, this research was conducted to show that these tests are not like apples and oranges and that score comparisons might be legitimate across these batteries. As far as face validity is concerned, the two tests might seem to be designed for different purposes: TOEFL as a general proficiency indicator and IELTS as an ESP one. Moreover, the researcher's personal interviews with a number of subjects (20) indicated that the majority of the testees preferred IELTS to TOEFL, believing that IELTS was a fairer indicator of their proficiency. The favourite section of IELTS, according to the subjects, was the reading section (ESP aspect), while the least favourite one was reckoned to be the listening part. The subjects, in general, thought that they had performed better at IELTS.

The question is whether the ESP colouring of IELTS makes it distinguishable from a

general proficiency test. Criper and Davies (1988) have shown that in spite of the intention of the designers of ELTS to create a multi-factorial structure test, the internal structure is in favour of a uni-factorial one. That is to say, general proficiency (whatever one may call it) is a better predictor of ELTS overall score. Alderson's (1993) trial study on IELTS also supports this idea. Although moderate correlations (0.51-0.67) between IELTS and TOEFL subtests reported in appendix I indicate that perhaps each test is testing something different- or rather, say in a different way- the factor analysis (table 7) indicates the dominance of a primary factor on which all the subtests of TOEFL and IELTS loaded and of which TOEFL listening comprehension loaded highest. This factor may well be interpreted as a general listening ability. The second factor, where MCHE's structure and reading comprehension loaded highest, could be interpreted as a general ability of reading comprehension and structure recognition. It may follow then that both TOEFL and IELTS acted unifactorially for the subjects concerned here. This is in accordance with previous research findings (Swinton and Powers 1980:15) that TOEFL acted unifactorially for less proficient groups. The TOEFL total mean score in this study is 468 which is far less than the average mean score for Farsi speakers (504) reported by ETS (1992).

The above discussions may lead us to the conclusion that IELTS and TOEFL share similar internal structure and may thus provide similar information of our testees' language ability. This allows us to do score comparisons across these tests in a rather meaningful way.

The results shown in table 4 are in accordance with most universities' expectations of the performance of non-native speakers on these two tests (see language proficiency requirement section of most UK and Australian Postgraduate Prospectus booklets). Score comparisons in table 4 indicate that a score of 6 on IELTS is equated with a score of 550 on TOEFL (the minimum requirement for allowing non-native speakers to enter into a non-linguistics department), while a score of 6.5 on IELTS is roughly equated with 600 on TOEFL (the minimum requirement for entering into a linguistics department). The comparisons in table 5, however, violate this equation. While changes in the less proficient subjects do not much affect the equation of the two scores, the changes in the scores for more proficient subjects (above 70 on MCHE scale) affect the equation in a meaningful way. Candidates who might have been accepted into a programme of study on their TOEFL score (600) would probably be rejected had their IELTS score (6) been taken into consideration. A closer comparison between Group A scores and Group B scores may suggest that subjects with approximately the same language ability performed differently in the two tests. Table 5 figures imply that subjects' (Group B) familiarity with the IELTS sample test had a slight improvement effect on the overall IELTS band score. They also imply that TOEFL preparation courses had a much higher improvement effect on the total TOEFL score. The effect is more striking for more proficient subjects.

Since score comparisons between Group A IELTS scores and Group B IELTS scores do not show much difference but the same comparisons between the two groups' TOEFL scores do show considerable difference, it may be concluded that TOEFL preparation courses had positive effect on the subjects' total TOEFL score. The overall MANOVA test shows that the effect of the factor "test" was significant. Moreover, the Tukey test suggests that the difference between the subjects' scores on TOEFL and IELTS was significant. It also implies that the subjects scored

significantly higher in TOEFL. This is in sharp contrast with what the subjects had earlier expressed in their interviews. Perhaps subjects' familiarity with the TOEFL format and their preparation courses were the main causes of this difference.

The Correlations reported in table 2 are not within one's expectation of the behaviour of similar LP tests. However, those reported in table 3 are well within one's expectation of the behaviour of LP tests. The difference might be due to the fact that scores reported here were gathered from different administrations of LP tests which might not have been equated to one another. So the difference might reflect the tests' unequated forms. It might also be due to the lower language ability of Group A subjects. Perhaps Group A subjects performed equally low at the two tests.

5. Conclusion

In this research we were looking for the justification of score comparisons across TOEFL and IELTS. We argued that since the internal structures of the two tests seem to be similar, tapping the same general proficiency factor, the tests may be comparable. It followed that score comparisons across the two test batteries are possible. The results of the comparisons suggested that although score comparisons across the two tests are possible, they might be affected by various factors across time. Factors such as test methods, subjects' familiarity with the test, LP preparation courses, and subjects' proficiency level might affect the score comparisons. This research was limited in scope to one native language only. Perhaps including the wide range of audience which these tests are addressing in the analysis would level the differences in score comparisons. Nevertheless, institutions using these test results should be cautioned about the relative comparability value of the test scores and should allow space for possible compromise of the band levels attached to the test scores. In short, score comparisons across LP tests are justified provided that possible extraneous factors affecting test scores are also taken into account.

Acknowledgements

I am grateful to Eric Glendinning for his comments on an earlier draft of this paper and to Dan Robertson for his help in doing the Tukey test. I alone am responsible for any mistakes.

References

Alderson C. 1978. 'A study of the cloze procedure with native and nonnative speakers of English'. Unpublished PhD dissertation. University of Edinburgh.

Alderson C. 1993. 'The relationship between grammar and reading in an English for academic purposes test battery' in D. Douglas and C. Chapelle (eds.) A New Decade of Language Testing Research. Alexandria, Virginia: TESOL. 203-219.

Bachman L. 1982. 'The trait structure of cloze test scores'. TESOL Quarterly 16,1: 61-70.

Bachman L. 1990. Fundamental Considerations in Language Testing. Oxford: OUP.

Bachman L. and A. Palmer. 1979. 'Convergent and discriminant validation of oral

language proficiency tests' in R. Silverstein (ed.) Proceedings of the Third International Conference on Frontiers in Language Proficiency and Dominance Testing. Carbondale, Ill.: Department of Linguistics, Southern Illinois University. 53-62.

Bachman L. and A. Palmer. 1981. 'The construct validation of the FSI oral interview'. Language Learning 31,1: 67-86.

Bachman L. and A. Palmer. 1982. 'The construct validation of some components of communicative proficiency'. TESOL Quarterly 16,4: 449-65.

Bachman L., F. Davidson and J. Foulks. 1993. 'A comparison of the abilities measured by the Cambridge and Educational Testing Service EFL test batteries' in D. Douglas and C. Chapelle (eds.) A New Decade of Language Testing Research. Alexandria, Virginia: TESOL. 25-46.

Bachman L., F. Davidson, K. Ryan, and I-C. Choie. Forthcoming. The Cambridge-TOEFL Comparability Study. Final Report. Cambridge: UCLES.

Campbell D. and D. Fiske. 1959. 'Convergent and discriminant validation by the multitrait-multimethod matrix'. Psychological Bulletin 56: 81-105.

Carroll J. 1978. English Language Testing Service. Specifications. London: The British Council.

Chapelle C. and R. Abraham. 1990. 'Cloze method: what difference does it make?' Language Testing 7,2: 121-146.

Criper C. and A. Davies. 1988. ELTS Validation Project Report 1. Hertford: The British Council and UCLES.

ETS. 1992. Bulletin of Information for TOEFL/TWE and TSE, 1992-93. Princeton, NJ.: Author.

Farhady H. 1980. 'Justification, development, and validation of functional language testing'. Unpublished PhD dissertation. University of California Los Angeles.

Lewkowicz J. 1983. 'Method Effect in Testing Reading Comprehension: a Case of Three Methods'. Unpublished MA dissertation. University of Lancaster.

Munby J. 1978. Communicative Syllabus Design. Cambridge: CUP.

Peirce B. 1992. 'Demystifying the TOEFL reading test'. TESOL Quarterly 26/4: 665-691.

Pike L. 1979. An Evaluation of Alternative Item Formats for Testing English as a Foreign Language. TOEFL research report 2. Princeton, NJ.: ETS.

Shohamy E. 1984. 'Does the testing method make a difference? The case of reading comprehension'. Language Testing 1,2: 147-70.

Stevenson D. 1987. 'Test of English as a Foreign Language' in C. Alderson, K. Krahnke, and C. Stanfield (eds.) 1987:79-81. Reviews of English Language

BEST COPY AVAILABLE

Proficiency Tests. Washington D.C.: TESOL.

Swinton S. and D. Powers. 1980. Factor Analysis of the TOEFL for Several Language Groups. Princeton, NJ.: ETS.

17

Appendix 1
Correlational Matrix Among Subsections of the Three Tests

	MLC	MST	MVOC	MRC	IRC	IWR	ILC	ISP	TLC	TST	TRC
MLC	1.0000										
MST	.4884**	1.0000									
MVOC	.4663**	.4055**	1.0000								
MRC	.4277**	.4709**	.4486**	1.0000							
IRC	.4584**	.3333**	.3278**	.1971	1.0000						
IWR	.4178**	.2991*	.2445*	.2524*	.5267**	1.0000					
ILC	.4678**	.3245**	.2477*	.2519*	.5473**	.4675**	1.0000				
ISP	.3508**	.2307*	.2869*	.2446*	.2897*	.4256**	.4823**	1.0000			
TLC	.5445**	.4181**	.3637**	.2076	.5211**	.4637**	.5619**	.4408**	1.0000		
TST	.4710**	.5593**	.4392**	.4064**	.5329**	.5069**	.4898**	.3944**	.6870**	1.0000	
TRC	.4468**	.5433**	.4827**	.3314**	.6072**	.4659**	.5044**	.3816**	.6877**	.7637**	1.00

N of cases: 103 1-tailed Signif: * - .01 ** - .001

M = MCHE, I = IELTS, T = TOEFL, LC = Listening Comprehension, ST = Structure, VOC = Vocabulary, RC = Reading Comprehension, WR = Writing, SP = Speaking

BEST COPY AVAILABLE